Available online at www.sciencedirect.com

SciVerse ScienceDirect

Genomics Proteomics Bioinformatics 10 (2012) 217–225

GENOMICS PROTEOMICS & BIOINFORMATICS

www.elsevier.com/locate/gpb

Original Research

Homepeptide Repeats: Implications for Protein Structure, Function and Evolution

Muthukumarasamy Uthayakumar [1,#], Bowdadu Benazir [1,#], Sanjeev Patra [1], Marthandan Kirti Vaishnavi [1], Manickam Gurusaran [1], Kanagarajan Sureka [2], Jeyaraman Jeyakanthan [2], Kanagaraj Sekar [1,*]

[1] *Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore 560012, India*
[2] *Department of Bioinformatics, Alagappa University, Karaikudi 630003, India*

Received 18 November 2011; revised 3 April 2012; accepted 19 April 2012
Available online 4 August 2012

Abstract

Analysis of protein sequences from *Mycobacterium tuberculosis* H37Rv (*Mtb* H37Rv) was performed to identify homopeptide repeat-containing proteins (HRCPs). Functional annotation of the HRCPs showed that they are preferentially involved in cellular metabolism. Furthermore, these homopeptide repeats might play some specific roles in protein–protein interaction. Repeat length differences among Bacteria, Archaea and Eukaryotes were calculated in order to identify the conservation of the repeats in these divergent kingdoms. From the results, it was evident that these repeats have a higher degree of conservation in Bacteria and Archaea than in Eukaryotes. In addition, there seems to be a direct correlation between the repeat length difference and the degree of divergence between the species. Our study supports the hypothesis that the presence of homopeptide repeats influences the rate of evolution of the protein sequences in which they are embedded. Thus, homopeptide repeat may have structural, functional and evolutionary implications on proteins.

Keywords: Homopeptide repeats; Disordered regions; Replication slippage; Protein domains; Rate of evolution

Introduction

Amino acid repeats are frequently found in protein sequences and are generated by repetitive elements in the genome like long terminal repeats (LTRs) and non-LTRs [1]. These repeats can be further classified into three distinct groups – homopeptide, dipeptide and sequence repeats, based on the number of amino acid residues repeated in the protein sequence [2]. Homopeptide repeats/single amino acid repeats are strings of a single amino acid residue occurring two or more times directly adjacent to each other. A dipeptide repeat is a pair of non-identical amino acid residues tandemly repeated in a linear sequence. Finally, a sequence repeat is an amino acid motif (made up of different combinations of amino acids) that is repeated several times in the protein sequence.

A homopeptide repeat is believed to have arisen either from sequential expansion of short codon repeat through replication slippage or from accumulation of point mutations in the coding sequences [2]. Several studies have shown that accumulation of guanine or cytosine at the third position of every codon in the genome is the fundamental cause for repeat expansion through replication slippage [3–5]. Homopeptide repeats are often embedded in low complexity regions, which also include interrupted and non-tandem repeats [6]. Hydrophobic residues are underrepresented in the repeats with the exception of leucine (L) which is over-represented in bacterial repeats and occurs in large numbers in signal peptides from humans [7]. Zhang et al. examined the location of homopeptide repeats in the protein sequences from *Arabidopsis thaliana* and *Oryza sativa* and identified a negative correlation

# Equal contribution.
* Corresponding author.
E-mail: sekar@physics.iisc.ernet.in (Sekar K).

http://dx.doi.org/10.1016/j.gpb.2012.04.001

between the amino acid distribution, their position and their functions [8]. The positional bias of some repeats to the N-terminal regions implies that these regions are more active in generating repeat sequences.

A recent work by Faux et al. revealed that all repeats (irrespective of the type of amino acid residue involved) play an important role in the processes that require the assembly of large multi-protein complexes [9]. Bjorklund et al. found that tandem repeats have a variety of binding properties and are involved in protein–protein interactions as well as binding to ligands such as DNA and RNA [10]. Moreover, homopeptide repeat that mediates protein–protein interactions can also facilitate network evolution [11]. Many parasitic organisms (Eukaryote or Bacteria) possess surface antigens that are made up of amino acid repeats. When forming an interface between host and pathogen, these repetitive proteins may act as virulence factors and get involved in immune invasion and cytoadherence [12].

Homopeptide repeats are also actively associated with the development of many diseases. For example, poly- glutamine (poly-Q) and poly-alanine (poly-A) repeats are involved in the development of Huntington disease and oculopharyngeal muscular dystrophy (OPMD), respectively [13]. Similarly, poly-glycine (poly-G) repeats play an important role in protein targeting [14] and poly-R repeats are involved in binding of viral proteins to RNA [15,16].

In *Mycobacterium tuberculosis* H37Rv (*Mtb* H37Rv), 21% of the proteins are hypothetical ones whose function has not yet been determined (results not shown). About 10% of the coding capacity has been dedicated to two Mycobacterium-specific protein families of unknown function, namely the Proline-Glutamate (PE) and Proline-Proline-Glutamate (PPE) families. Proteins belonging to these two families have conserved proline (P) and Q residues in their N-termini [17]. There are many structural and functional studies on Mycobacterium gene products as a whole, but still no sufficient data is available about the large number of homopeptide repeats present in the proteins.

Although homopeptide repeats have been studied and characterized in many other proteomes, information is unavailable about their structure, function and evolution particularly about those present in prokaryotes. The present study provides an overall picture of the homopeptide repeats present in *Mtb H37Rv* proteome and the repeats were analyzed in relation to their implication in protein structure, function and evolution.

Results and discussion

During the present analysis, homopeptide repeat-containing proteins (HRCPs) are found to be 289 (a total of 310 repeats), which is about 7.7% of the *Mtb* H37Rv protein sequences (see Materials and Methods for details). The proportion of the repeat-containing proteins in *Mtb* H37Rv is similar to that of yeast; whose protein sequences contain 7.6% homopeptide repeats [18].

Distributional gradients of homopeptide repeats in Mtb H37Rv

Faux et al. previously reported that in prokaryotes, homopeptide repeats are commonly made up of serine (S), G, A and P, while in eukaryotes, the amino acids Q, aspargine (N), A, S and G are often seen in homopeptide repeats [9]. **Figure 1** illustrates the rate of occurrence of homopeptide repeats based on their repeat frequencies. From the figure, it is evident that four distinct classes of homopeptide repeats containing A, G, arginine (R) or P are abundant in *Mtb* H37Rv. These results are consistent with that of Faux et al. Green and Wang proposed that the repeats formed by hydrophilic amino acids are abundant in the protein sequence databases [19]. However, it is interesting to note that a majority of the homopeptide repeats observed in the present study are composed of hydrophobic amino acid residues which in turn could affect the overall hydrophobicity of the protein [20]. In addition, Marcotte and his co-workers proposed that R repeats are highly depleted in protein sequences [21]. Nevertheless, in the *Mtb* H37Rv genome, R repeats constitute about 8% of the repeats present (Figure 1). Arginine being a charged amino acid may play an important role in protein–protein interactions.

Functional annotation of HRCPs in Mtb H37Rv

Experimental evidence suggests that HRCPs have biased functions [8,9,18]. Functional annotation of the HRCPs is carried out using Yeast Protein Database (YPD) to evaluate their abundance in different classes of proteins. Results showed that, in yeast, homopeptide repeats are overrepresented in transcription factors, protein kinases and transporter proteins [9]. These data are consistent with

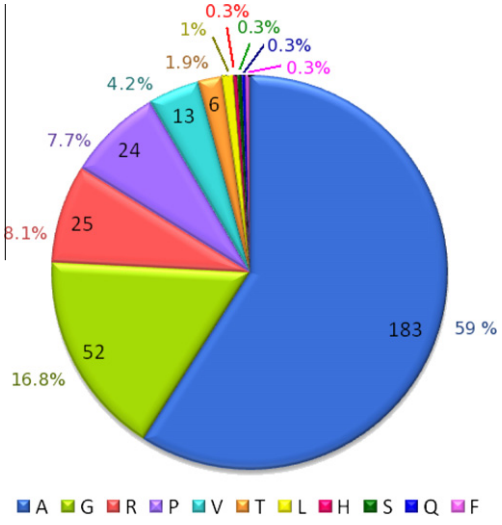

Figure 1 Total number and percentage of homopeptide repeats in *Mtb* H37Rv

A pie chart describing the total number and the percentage of the homopeptide repeats. The residues A (blue), G (green), R (red) and P (violet) occur most often among the homopeptide repeats.

studies by Gerber et al. indicating that the insertion of either poly-Q or poly-P tracts enhances the transcriptional activation of the GAL4/VP16 fusion construct in yeast [22]. Furthermore, analysis on the HRCPs from *Homo sapiens*, rodents, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Oryza sativa* and *Arabidopsis thaliana* all showed similar observations [9,23].

In *Mtb* H37Rv, a total of 289 HRCPs exhibit have functional annotations. However, and the functions of these HRCPs are not well established. 58% (168 out of 289) of the HRCPs were assigned to four protein families – polymorphic GC-rich repetitive sequence (PGRS) proteins, PPE proteins, PE proteins and hypothetical proteins (**Figure 2**). However, about 22% (65 HRCPs) of the HRCPs are involved in cellular metabolism and 12% (34 HRCPs) of the proteins participated in transport and signaling process (grouped as "Transport" in Figure 2). Only 3% (10 HRCPs) of HRCPs belong to either transcription or translation apparatus, indicating that in *Mtb* H37Rv, repeats do not exhibit bias towards transcriptional or translational proteins. This may be due to the fact that bacteria have simple intracellular mechanisms. Thus, from the present study, it appears that the selective enrichment of homopeptide repeats towards transcription factors and signaling proteins does not extend to *Mtb* H37Rv protein sequences.

Scanning for domains containing homopeptide repeats

Bjorklund et al. observed that the homopeptide repeats were present within protein domains of the same family [11]. The homopeptide repeats present in *Mtb* H37Rv within the protein domain along with their Pfam identification numbers are given in Table S1. In *Mtb* H37Rv, 65% of the HRCPs contain homopeptide repeats within their functional domains and about 28% of the total repeats are present within the synthase, transferase and dehydrogenase domains (Table S1). To deduce the functions of these repeats in the domains, the three-dimensional (3D) structures of the HRCPs were given as input to the BSDD server [24]. The server identified two distinct domains, including peptide (small fragments) containing arginine repeats within the RNA-binding domain, and biotin carboxylase containing glycine within ATP-binding domain.

Identification of homopeptide repeats in 3D structure

Previous studies reported that compared to the protein sequence database, there is a deficiency of simple sequence repeats in PDB [25,26]. Homopeptide repeats encoded by iterations of a single codon likely result from strand slippage and substantial increase in multiple codons within the homopeptide coding regions which makes the homopeptides important for the structure of the proteins containing them [27]. The 3D structures of only five HRCPs from *Mtb* H37Rv are available in PDB (**Table 1**). Among these five structures, two repeats interact with a metal ion (histidine tracks in zinc and ferric uptake regulator protein with PDB ID 2O03). Poly-A repeats are well-known for their helical structure [28] and in *Mycobacterial* protein sequences, poly-A stretches form helical structures too.

As most of the HRCPs in *Mtb* H37Rv are yet to be crystallized, an extensive search for the homopeptide repeats in PDB was carried out using the BSDD server [24]. Simultaneously the corresponding secondary structures were analyzed using the STRIDE software [29]. Serine, glycine and arginine repeats have high probability of forming β turn conformation, while leucine is involved in α helical conformation. Protein domains tend to exhibit stable yet flexible conformations when present in the short linker groups. Thus, homopeptide expansions are tolerated when they occur in the linker regions. In *Mtb* H37Rv, histidine repeat in the linker region is found in serine/threonine pro-

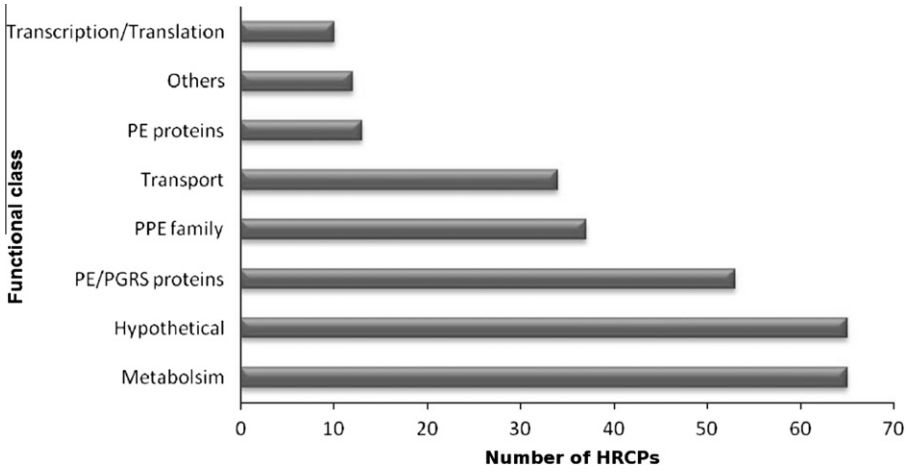

Figure 2 Functional annotation of HRCPs in *Mtb* H37Rv
The homopeptide repeats were analyzed and categorized into eight classes. The number of repeat containing proteins has been provided on the *X*-axis and their corresponding functional class is provided on the *Y*-axis. The highest number of repeat containing proteins is the metabolic and hypothetical proteins.

Table 1 Structural homopeptide repeats present in *Mtb* H37Rv protein sequences

PDB ID	Protein description	Repeat type	Secondary structure
2O03	Zinc uptake regulator protein	HHHHH	β-sheets
1RWI	Serine/threonine protein kinase D	HHHHHH	β-turns
1U5H	Citrate lyase beta subunit	AAAAA	α-helix
2BPQ	Anthranilate phosphoribosyl transferase	AAAAA	α-helix
3EKL	Fructose 1,6-bis phosphate aldolase	TTTTT	β-hairpins

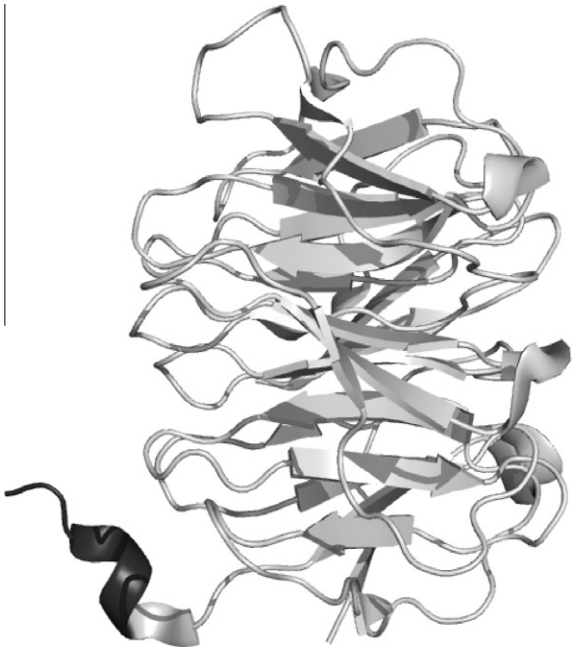

Figure 3 The poly-histidine tract occurs in the linker regions of protein kinase PknD
The poly-H track (residues 265–270) was revealed in the linker regions of Ser/Thr protein kinase PknD (PDB ID: 1RWI). Poly-H was shown in black.

tein kinase PknD. **Figure 3** shows the extracellular domain of *Mtb* PknD (PDB ID: 1RWI) with histidine repeats present in the flexible linker regions.

Repeats present in the disordered regions

Disordered regions are parts within the protein molecule that do not fold into a stable secondary structure. These regions may vary in size [30,31] and are involved in many biological processes such as regulation, signaling and cell cycle control [32]. To check if homopeptide repeats are disordered, the HRCPs were given as input to the RONN software. An amino acid residue is determined as disordered if the average probability of disorder (also known as RONN score) is greater than 0.5 [33]. As shown in **Tables 2** and S2, six homopeptide repeat types made up of R, P, G, H, S or Q have a RONN score greater than 0.5, suggesting that these repeats are disordered, hence lack a definite tertiary structure. In addition, two repeat types made up of A or T demonstrate an average score of 0.47 (can be considered as 0.5). Thus, it is reasonable to conclude that 90% of the homopeptide repeats found in *Mtb* H37Rv are disordered since 9 out of 11 repeat types observed have a RONN score greater than or equal to 0.5 (Tables 2 and S2).

Table 2 Average RONN score for the homopeptide repeats in the functionally annotated HRCPs

Repeat type	Number of repeats	Average RONN score
Alanine (A)	78	0.49
Arginine (R)	11	0.64
Proline (P)	12	0.76
Glycine (G)	6	0.52
Histidine (H)	1	0.51
Valine (V)	8	0.33
Threonine (T)	3	0.47
Leucine (L)	2	0.32
Serine (S)	1	0.52
Glutamine (Q)	1	0.91
Phenylalanine (F)	1	0.19

Amino acid usage, frequency of amino acid repeats and its usage

We next tested whether there is a linear correlation between the frequency of amino acid repeats and the usage of the amino acid in the protein sequences in *Mtb* H37Rv. The compositional analysis of all the protein sequences shows that A, G, L and V are found in abundance (**Figure 4**). In the present study, valine is not further considered since the overall count of valine rich repeats is low. There is a significant difference in the amino acid usage and the frequency of the corresponding homopeptide repeat present in all the *Mtb* H37Rv protein sequences. The percentage of the homopeptide repeats is plotted against the composition of the corresponding amino acids in the protein sequences in Figure 4. From this figure, it is evident that only four amino acid residues (A, G, P and R) have a higher tendency to form repeat sequences. Although the other amino acid residues (like L) are present relatively high, they possess a very low propensity to form homopeptide repeats. These results are consistent with the hypothesis proposed by Zhang et al. that poly-G repeats are solely responsible for the high glycine content in *Oryza sativa* and *Arabidopsis thaliana* [8]. These data suggest that there is no linear relationship between the frequency of the amino acid repeats and amino acid composition [1]. Thus, it is possible that other than the amino acid content, many factors are responsible for the occurrence and frequencies of the homopeptide repeats.

Difference in repeat size

Comparative studies on repeat size difference between species have dealt only with a small number of genes. It is therefore unclear whether the rapid change reported previ-

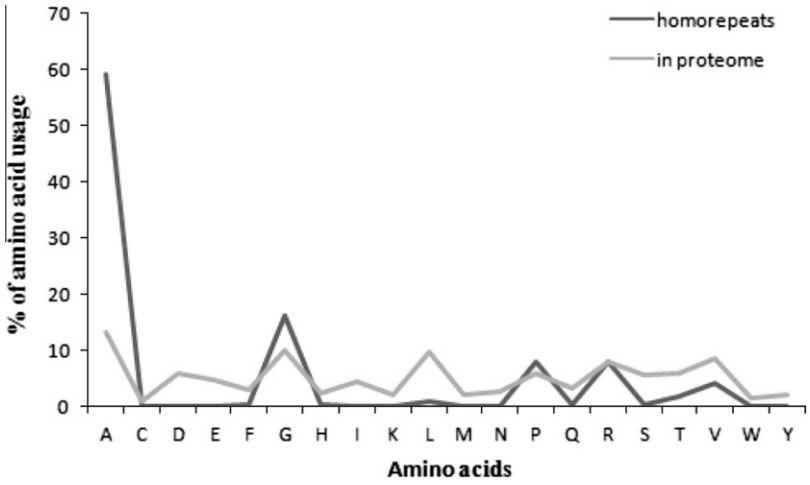

Figure 4 Comparison of amino acid usage in homopeptide repeats and proteome in *Mtb* H37Rv
The amino acid usage (%) in both homopeptide repeats (homerepeats) and in the protein sequences (in proteome) was provided here. The amino acid residues A, G, P and R exhibit peak usage in both homopeptide repeats and the total protein sequences.

ously is representative of homopeptide repeats in general or whether conserved homopeptide repeats also exist [18]. To address this question, we compared the homopeptide repeats from four homologous species. Repeat size difference between homologous proteins indicates a degree of repeat conservation. The size difference ranges between 0 and 1, where 0 represents complete conservation of the repeats and 1 implies complete absence of the repeats between two homologous sequences [34].

We examined the HRCPs from *Mtb* H37Rv and their putative homologs from three other organisms including *Escherichia coli* (Bacteria), *Sulfolobus acidocaldarius* DSM639 (Archaea) and *Homo sapiens* (Eukaryote). We found that the repeat sequences followed a regular pattern of conservation and divergence between the four species (**Figure 5**). The repeat size difference between *Mtb* H37Rv and *E. coli* is low, compared to that between *Mtb* H37Rv and *S. acidocaldarius DSM 639* or *H. sapiens* (Table S3 and Figure 5). It is interesting to note that the repeats are more conserved in *Mtb* H37Rv and *E. coli* and least in *H. sapiens*, demonstrating that there exists a direct relationship between the repeat size difference and the order of divergence.

Evolutionary rate analysis

Low-complexity regions that include homopeptide repeats evolve much faster than the rest of the protein sequence [35] and homopeptide repeats are present within functionally and structurally more evolutionarily active regions [36]. It was hypothesized that repeat expansion and contraction may provide a mechanism for rapid morphological evolutionary changes [37] and that this expansion and contraction may be facilitated by replicative slippage [38,12]. An evolutionary analysis of 2838 open reading frames from three Saccharomyces species showed that fast evolving low-complexity sequences outnumbered conserved sequences by a ratio of 10–1 [35]. Previous studies on Hox proteins by Casillas et al. showed that the long homopeptides are present all along the protein except in the highly conserved regions [36]. These repeat regions are the origin of most of the indels and thus are responsible for the high amino acid evolution of Hox proteins.

Previous studies show that homopeptide repeats have an influence on the evolutionary rate of the protein sequences in which they are embedded [25,37,39]. To see whether this holds true in Mycobacteria, orthologous proteins from

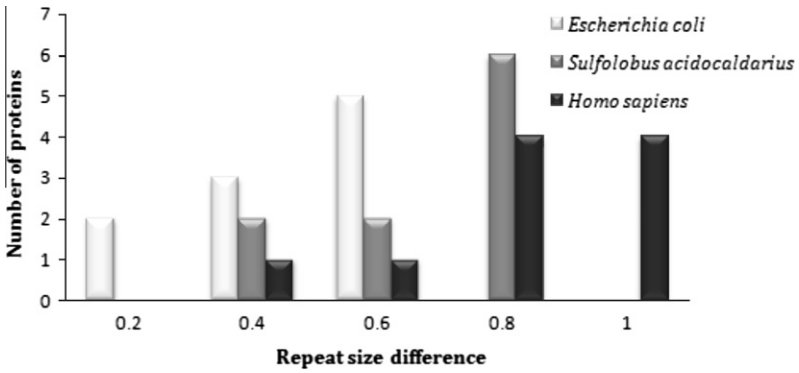

Figure 5 Repeat size differences between proteins from *Mtb* H37Rv, *E. coli, S. acidocaldarius* and *H. sapiens*
The repeat size difference between *Mtb* H37Rv and *E. coli* is low, compared to that between *Mtb* H37Rv and *S. acidocaldarius DSM 639* or *H. sapiens.*

Table 3 Homopeptide repeat frequency in the seven orthologous Mycobacterium

Homopeptide repeat type	Homopeptide repeat frequency in the Mycobacterial strains						
	Mtb H37Rv	*M. marinum*	*M.* MCS	*M. smegmatis*	*M. ulcerans*	*M. gilvum*	*M. vanbaalenii*
Alanine	59.35	52.9	47.16	56.28	63.97	47.34	50.15
Glycine	16.12	23.0	6.9	7.6	11.49	7.4	8.9
Proline	8.0	7.9	20.0	13.66	7.1	19.78	17.23
Arginine	8	4.3	8.5	5.76	4.65	10.6	9.8
Valine	4.19	3.19	4.12	3	3.41	4.5	3.6
Threonine	1.9	2.7	3.8	5.19	2.17	3.5	3.38
Leucine	0.96	2.2	4.76	2.7	2.79	2.4	2.15
Serine	0.32	0.6	1.58	2.45	1.8	1.76	1.2
Glutamine	0.32	0.4	0.6	1.09	0.3	0.3	0.3
Histidine	0.32	0.2	0.3	0.2	0.3	0.3	0.6
Phenylalanine	0.32	0	0.3	0.2	0	0.3	0.3
Aspartic acid	0	1.3	1.2	1.6	1.2	1.4	2.15
Isoleucine	0	0.2	0	0	0.3	0	0
Glutamic acid	0	0.2	0	0	0	0	0
Lysine	0	0.2	0	0	0.3	0	0

seven *Mycobacterial* species are included in the present study. The occurrence of homopeptide repeats is consistent among the seven *Mycobacterial* species, however, their frequencies are not similar (**Table 3**). Twenty orthologous proteins contained several homopeptide repeats in common. Multiple sequence alignment for all 20 orthologous proteins showed that the homopeptide repeats observed in these proteins are conserved among the seven Mycobacterial species. An example of the alignment is provided in **Figure 6** for mmpS3. The repeats in the protein sequences are conserved at different levels, which is evident from the expansion and contraction of the repeats in the seven species, *e.g.*, the proline tract found in mmpS3 shown in Figure 6. Furthermore, we constructed the phylogenetic tree based on the multiple sequence alignment of all 20 proteins. The evolutionary rates with repeat blocks, E(R), and without repeat blocks, E(A), were calculated for all 20 proteins using PROTDIST and FITCH programs [40] (**Table 4**). Paired t-test analysis showed that the evolutionary rates of all 20 proteins with repeat blocks are significantly different (at 0.05 level) from that of the proteins without repeat blocks. In addition, a binomial testing was performed for all 20 proteins and the resulting Z-score (4.504) supported the significant difference (at 0.01 level). Thus, E(R) and E(A) result from different populations and the mean difference observed between them is not a consequence of coincidence or random sampling. Taken together, these data indicate that the differences observed in the experiment are not a consequence of coincidence or random sampling. Therefore, it is reasonable to conclude that proteins with homopeptide repeats E(R) evolve faster than the proteins without repeats E(A).

Table 4 Rates of proteins with repeat and without repeat blocks

Protein	Repeat	E(R)	E(A)
murG	RRRRR	0.69	0.68
Lppr	TTTTT	3.54	3.31
Ask	VVVVV	48.13	32.3
accA1	VVVVV	0.86	0.67
dinF	VVVVV	0.70	0.69
mmpS3	PPPPPP	1.05	0.02
nusA	PPPPP	0.30	0.02
pknE	PPPPPP	2.33	0.82
mutB	AAAAAA	0.30	0.2
trpC	AAAAAA	10.11	0.3
1pqI	AAAAA	1.06	1.00
thiE	AAAAA	0.86	0.76
menE	AAAAA	27.74	1.02
menC	AAAAA	0.78	0.76
metS	AAAAA	22.23	1.31
rhlE	RRRRRRRRR	0.71	0.63
MCE1C	PPPPPP	1.04	1.00
MCE4F	PPPPPP	0.87	0.8
Mycosin	PPPPPP	2.03	2.00
FadE19	TTTTTT	0.40	0.35

Note: E(R), evolutionary rate of proteins with repeat blocks; E(A), evolutionary rate of proteins without repeat blocks.

Conclusion

The present analysis shows that the homopeptide repeats are common in *Mtb* H37Rv and four major classes are identified, namely, poly-A, poly-G, poly-R and poly-P con-

```
gi|120404533|ref|YP_954362.1|   118 -PAPPVQDEFTTTEPP---PPP----TTEPPPPPP--ETVTVTEPPPPPP 157
gi|145223538|ref|YP_001134216.  108 -QAPPVQDEILTTTPP---PPPPPPPTTEPPPPPPPPETVTVTQPPPPPP 153
gi|108800262|ref|YP_640459.1|   135 TPPPPVQDEITTTTPP---PPPPPPTTEPPPPPPP-ETVTVTEPPPPPP 180
gi|118472224|ref|YP_888542.1|   132 SSAPPVQDEITTTTPPPVEPPPPPPPSEEAPPPPPP-ETVTITEPPPAPA 180
gi|118618840|ref|YP_907172.1|   147 SSVPPVQDEITTTKPPPPPPPPPPPPSTEIPTATET-QTVTVTPPPPPPP 195
gi|183983233|ref|YP_001851524.  147 SSVPPVQDEITTTKPPPPPPPPPPPPSTEIPTATET-QTVTVTPPPPPPP 195
gi|15609335|ref|NP_216714.1|    136 SSAPPVQDEITTTKPAPPPPPAPPPTTEIPTATET-QTVTVTPPPPPP- 183
```

Figure 6 Sequence alignment of orthologous mmpS3 protein from seven Mycobacterium
The sequence highlighted within the box represents the homopeptide repeat block. The start and the end positions of the protein sequences was provided before and after the sequences.

taining repeats. Structural and functional analyses show that each class of homopeptide repeats adopt a distinct secondary structure and are mainly found in proteins involved in cellular metabolism, transport and signaling. Analysis of the homopeptide repeat frequency and the amino acid usage suggests that the amino acid content is not the only factor responsible for changes in homopeptide repeat frequencies. Finally, homopeptide repeats could have a profound effect on the evolutionary rate of the protein containing it. Thus, it can be concluded that homopeptide repeats play a vital role in the structure, function and evolutionary rate of the homopeptide repeat containing proteins.

Materials and methods

Identification and distribution of homopeptide repeats in Mtb H37Rv

All the protein sequences from *Mtb* H37Rv (3988 protein sequences) were downloaded from the local FTP site maintained at the Bioinformatics centre, Indian Institute of Science. Locally developed PERL scripts were used to detect the homopeptide repeats in the protein sequences. Homopeptide repeats defined here refer to continuous occurrence of single amino acid residues for five or more times. The cut-off size of five residues was chosen because of its significantly low probability of occurrence by chance [41].

Repeat frequency = Number of X repeat/total number of homopeptide repeats present in the protein sequences, where X represents any amino acid residue.

Functional annotation of HRCPs

Functional annotation of HRCPs was done using NCBI BioSystems [42] and COG databases [43]. NCBI BioSystems database was used to detect the protein structures involved either in a biological or disease pathway. In the case of COG, the most straightforward application is the prediction of individual protein function using the COGNITOR program. For each HRCP, the NCBI RefSeq_ID (Reference Sequence IDentifier), description, homopeptide tract and function were recorded.

Scanning for domains containing homopeptide repeats

The HRCP sequences were searched against the Pfam database [44] to identify the functional domains present. Two programs on the PDBsum [45] server namely, LIGPLOT [46] and NUCPLOT [47], was used to analyze the molecular interactions. LIGPLOT gives the schematic representation of the interactions between the ligand and the amino acid residues, while NUCPLOT shows the protein-nucleic acid interactions. Here again, the RefSeq_ID, description, repeat type, domain name and Pfam identification were recorded for each homopeptide repeat present within the protein domain.

Identification of homopeptide repeats in 3D structure

To identify the secondary and tertiary structures of the HRCPs, the PDB archive was searched with the standalone versions of BLASTP [48] and STRIDE softwares [29].

Identification of homopeptides present in the disordered regions of the proteins

Functionally annotated HRCP sequences were probed for disordered regions. RONN [33] software was used to detect the degree of disorder in the HRCP sequences.

Calculation of the amino acid usage

The amino acid usage and the homopeptide repeat fraction (number of homopeptide repeats/total number of proteins) in *Mtb* H37Rv were calculated using locally developed PERL scripts. The amino acid frequency of all the protein sequences was calculated using the formula:

Amino acid frequency = Number of X residues in the protein sequence/total number of amino acid residues in the protein sequences, where X represents any amino acid residue.

The amino acid frequencies of all the protein sequences were compared with the homopeptide repeat fraction observed.

Calculation of the repeat size difference

To investigate the purifying selection pressure on homopeptides, putative homologous HRCPs from *Mtb* H37Rv, *E. coli* (Bacteria), *Sulfolobus acidocaldarius* DSM639 (Archaea) and *Homo sapiens* (Eukaryote) were detected using the BLASTP [47] search (against the nr database). The putative homologous sequences were aligned using ClustalW [49] and scanned for the homopeptide tract. The numbers of amino acids in homopeptide tracts of the three homologs were recorded.

Evolutionary rate analysis

HRCPs from *Mtb* H37Rv were given as query to the program BLASTP [48]. Using H37Rv as the template, 20 orthologous proteins were detected in six other Mtb species (*Mycobacterium marinum, Mycobacterium smegmatis* MC2155 strain, *Mycobacterium ulcerans* Agy99 strain, *Mycobacterium vanbaaleni* PYR-1 strain, *Mycobacterium sp.* MCS strain and *Mycobacterium gilvum* PYR – GCK strain). These orthologous protein sequences were aligned using ClustalW [49] and the alignments were scanned to identify the homopeptide repeats of at least five residues

long. The multiple sequence alignment of the proteins was provided as input to the PROTDIST program of the PHYLIP (the PHYLogeny Inference Package) software [40] for the construction of distant matrices (based on Jones-Taylor-Thornton distance model), which were then used to construct phylogenetic trees using the FITCH program in the same package. The sum of the branch lengths in the phylogenetic tree provided the evolutionary rates for the proteins with repeat sequences. Further, to investigate the influence of the repeats in the evolution of the protein, the repeat blocks were removed and the above mentioned procedure was applied to the protein sequences.

Authors' contributions

KS conceived and supervised the project, and critically analyzed the database. MUK, BB, SP, JJ and KS collected the data and performed various analyses. MG performed the statistical analysis for validation. BB drafted the manuscript. MKV critically reviewed the manuscript and participated in discussion, curation and validation. KS revised the manuscript. All authors read and approved the final manuscript.

Competing interest

The authors have no competing interest to declare.

Acknowledgements

The present work is fully supported and funded by the Department of Information Technology (DIT), Government of India (Grant No. MITO-088). The authors gratefully acknowledge the facilities offered by the Interactive Graphics Facility and the Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore.

Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gpb.2012.04.001.

References

[1] Depledge DP, Dalby AR. COPASAAR – a database for proteomic analysis of single amino acid repeats. BMC Bioinformatics 2005;6:196.
[2] Depledge DP, Lower RPJ, Smith DF. RepSeq – a database of amino acid repeats present in lower eukaryotic pathogens. BMC Bioinformatics 2007;8:112.
[3] Cocquet J, De Baere E, Caburet S, Veitia RA. Compositional biases and poly-A runs in humans. Genetics 2003;165:1613–7.
[4] Caburet S, Vaiman D, Veitia RA. A genomic basis for the evolution of vertebrate transcription factors containing amino acid runs. Genetics 2004;167:1813–20.
[5] Nakachi Y, Hayakawa T, Oota H, Sumiyama K, Wang L, Ueda S. Nucleotide compositional constraints on genomes generate alanine-, glcyine-, and proline-rich structures in transcription factors. Mol Biol Evol 1997;14:1042–9.
[6] Hancock JM, Worthley EA, Santibanez-Koref MF. A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice. Mol Biol Evol 2001;18:1014–23.
[7] Labaj PP, Leparc GG, Bardet AF, Kreil G, Kreil DP. Single amino acid repeats in signal peptides. FEBS J 2010;277:3147–57.
[8] Zhang L, Yu S, Cao Y, Wang J, Zuo K, Qin J, et al. Distributional gradient of amino acid repeats in plant proteins. Genome 2006;49:900–5.
[9] Faux NG, Bottomley SP, Lesk JA, Morrison JR, Banda MG, Whisstock JC. Functional insights from the distribution and role of homopeptide repeat containing proteins. Genome Res 2005;15:537–51.
[10] Bjorklund AK, Ekman D, Elofsson A. Expansion of protein domain repeats. PLoS Comput Biol 2006;2:114.
[11] Hancock JM, Simon M. Simple sequence repeats in proteins and their significance for network evolution. Gene 2005;345:113–8.
[12] Niklaus F, Nguyen-Ha TM, Adler J, Maser P. Surface antigens and potential virulence factors from parasites detected by comparative genomics of perfect amino acid repeats. Proteome Sci 2007;5:20.
[13] Fandrich M, Dobson CM. The behaviour of polyamino acids reveals an inverse side chain effect in amyloid structure formation. EMBO J 2002;21:5682–90.
[14] Inoue K, Keegstra K. A polyglycine stretch is necessary for proper targeting of the protein translocation channel precursor to the outer envelope membrane of chloroplasts. Plant J 2003;34:661–9.
[15] Calnan BJ, Tidor B, Biancalana S, Hudson D, Frankel AD. Arginine-mediated RNA recognition: the arginine fork. Science 1991;252:1167–71.
[16] Nam YS, Petrovic A, Jeong KS, Venkatesan S. Exchange of the basic domain of human immunodeficiency virus type 1 Rev for a polyarginine stretch expands the RNA binding specificity, and a minimal arginine cluster is required for optimal RRE RNA binding affinity, nuclear accumulation, and trans-activation. J Virol 2001;75:2957–71.
[17] Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 1998;393:537–44.
[18] Alba MM, Satibanez-Koref MF, Hancock JM. Amino acid reiterations in yeast are overrepresented in particular class of proteins and show evidence of a slippage-like mutational process. J Mol Evol 1999;49:789–97.
[19] Green H, Wang N. Codon reiteration and the evolution of proteins. Proc Natl Acad Sci U S A 1994;91:4298–302.
[20] Oma Y, Kino Y, Sasagawa N, Ishiura S. Intracellular localization of homopolymeric amino acid-containing proteins expressed in mammalian cells. J Biol Chem 2004;279:21217–22.
[21] Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. A census of protein repeats. J Mol Biol 1999;293:151–60.
[22] Gerber HP, Seipel K, Georgiev O, Hofferer M, Hug M, Rusconi S, et al. Transcriptional activation modulated by homo-polymeric glutamine and proline stretches. Science 1994;263:808–11.
[23] Alba MM, Guigo R. Comparative analysis of amino acid repeats in rodents and humans. Genome Res 2003;14:549–54.
[24] Selvarani P, Shanthi V, Rajesh CK, Saravanan S, Sekar K. BSDD: Biomolecules Segment Display Device – a web-based interactive display tool. Nucleic Acids Res 2004;32:W645–8.
[25] Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence database. Comput Chem 1993;17:149–63.
[26] Saqi M. An analysis of structural instance of low complexity sequence segments. Protein Eng 1995;8:1069–73.
[27] Karlin S, Burge C. Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. Proc Natl Acad Sci U S A 1996;93:1560–5.
[28] Rohl CA, Fiori W. Baldwin RL Alanine is helix-stabilizing in both template-nucleates and standard peptide helices. Proc Natl Acad Sci U S A 1999;96:3682–7.

[29] Heinig M, Frishman D. STRIDE: a web server for secondary structure assignment from the known atomic coordinates of proteins. Nucleic Acids Res 2004;32:502–4.
[30] Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 1999;293:321–31.
[31] Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci 2002;27:523–33.
[32] Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry 2002;41:6573–82.
[33] Yang ZR, Thomson R, McNeil P, Esnouf RM. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 2005;21:3369–75.
[34] Mularoni L, Veitia RA, Alba MM. Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. Genomics 2006;89:316–25.
[35] Romov PA, Li F, Lipke PN, Epstein SL, Qiu WG. Comparative genomics reveals long, evolutionarily conserved, low-complexity islands in yeast proteins. J Mol Evol 2006;63:415–25.
[36] Casillas S, Negre B, Barbadilla A, Ruiz A. Fast sequence evolution of Hox and Hox-derived genes in the genus *Drosophila*. BMC Evol Biol 2006;6:106.
[37] Fondon JW, Garnes HR. Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci U S A 2004;101:18058–63.
[38] Rauceo JM, Armond RD, Otoo H, Kahn PC, Klotz SA, Gaur NK, et al. Threonine-rich repeats increase fibronectin binding in the *Candida albicans* Adhesin als5p. Eukaryot Cell 2006;5:1664–73.
[39] Huntley M, Clark AG. Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. Mol Biol Evol 2007;24:2598–609.
[40] Felsenstein J. PHYLIP – phylogeny inference package (Version 3.2). Cladistics 1989;5:164–6.
[41] Karlin S. Statistical significance of sequence patterns in proteins. Curr Opin Struct Biol 1995;5:360–71.
[42] Lewis YG, Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, et al. The NCBI BioSystems database. Nucleic Acids Res 2009;38:D492–6.
[43] Tatusov RL, Galperin MY, Nattale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein function and evolution. Nucleic Acids Res 2000;1:33–6.
[44] Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, et al. The Pfam protein families database. Nucleic Acids Res 2008;36:D281–8.
[45] Laskowski RA. PDBsum: summaries and analyses of PDB structures. Nucleic Acids Res 2001;29:221–2.
[46] Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng 1995;8:127–34.
[47] Luscombe NM, Laskowski RA, Thornton JM. NUCPLOT: a program to generate schematic diagrams of protein-nucleic acid interactions. Nucleic Acids Res 1997;25:4940–5.
[48] Altschul SF, Gish W, Miller W, Myers W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
[49] Higgins D, Thompson J, Gibson T, Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–80.